SPEAKER-INDEPENDENT MALAY VOWEL RECOGNITION OF CHILDREN USING MULTI-LAYER PERCEPTRON

Hua Nong TING and Jasmy YUNUS

Department of Electronics, Faculty of Electrical Engineering,
University of Technology, Malaysia
tinghuanong@hotmail.com

ABSTRACT

Most of the speech recognitions are based on adult speech sounds. Less research is done in the recognition of children speech sounds. The speech of children is more dynamic and inconsistent if compared to adult's speech. This paper investigates the use of neural networks in recognizing 6 Malay vowels of Malay children in a speaker-independent manner. Multi-layer Perceptron with one hidden layer was used to recognize these vowels. The Multi-layer Perceptron was trained and tested with speech samples of Malay children with their ages between seven and ten years old. A single frame of cepstral coefficients were extracted around the vowel onset point using Linear Predictive Coding. The vowel length was examined from 5 ms to 70 ms. Experiments were conducted to determine the optimal vowel length as well as the number of cepstral coefficients.

1. INTRODUCTION

Most of the speech recognitions involve adult speech and less research is conducted to look into the children speech recognition. The speech characteristics of children are more dynamic and inconsistent if compared to adult's speech. The fundamental and formant frequencies of children are higher than adult's [1]. Besides that, children show higher intra and inter speaker acoustic variability than adult speech in spectral and temporal characteristics.

The speech signal is dynamic in nature. Thus, in order to assume that it is stationary, the speech signal needs to be examined in a short segment. For example, the vowels are examined in its stationary region, in a fixed short frame of 25.6 ms [2]. The frame length actually can be variable from 5 ms up to 40 ms. The paper describes the effect of the variability of the frame length on the vowel recognition accuracy.

Linear Predictive Coding (LPC) is one of the most popular techniques in extracting speech features. For most of the speech recognition system, the order of LPC is fixed at a constant value, such as 10 or 12 order at a sampling rate of 8 kHz. The order of LPC is dependent on the sampling rate [3]. Instead of just fixing the order of the LPC, the paper investigates the effect of different LPC order on the performance of the vowel recognition.

Multi-layer Perceptron (MLP) is used to classify the Malay vowels. The training error and the hidden neuron number of MLP are normally fixed for the training and testing [2,6]. The performance of MLP depends on the appropriate setting of hidden neuron number as well as the training error. The effects of the hidden neuron number as well as the variability of training error on the recognition accuracy are described.

The experimental data is briefed in Section II. The speech feature extraction of Linear Predictive Coding is described in Section III. Section IV explains the architecture of the Multi-layer Perceptron. The experimental results and discussions are provided in Section V. The paper is ended with conclusion in Section VI.

2. EXPERIMENTAL DATA

There are 24 pure Malay phonemes in Malay language. The phonemes consist of 18 consonants and 6 vowels. Besides that, there are three diphthongs in Malay language: /ai/, /au/ and /oi/. The list of Malay vowels is shown in Table 1, which is based on the studies conducted by Hassan [4] and Karim [5].

Table 1: List of Malay vowels

Tongue Position Tongue Height	Front	Center	Back
High	i		u
Mid-high	e	ə	o
Mid-low			
Low	a		

A speech database was collected from Malay children with their ages between seven and ten years old in the primary religious school of University of Technology, Malaysia. For training set, 40 Malay children were involved. The group was consisted of 20 Malay males and 20 Malay females. As for testing set, the speech samples were collected from another

0-7803-8560-8/04/$20.00©2004IEEE

group of 20 Malay children, who was comprised of 10 males and 10 females.

The six Malay vowels were extracted from the six Malay words: “Gajah”, “Leher”, “Selipar” (/Səlipar/), “Filem”(/Filəm/), “Yoyo” and “Sudu”. Two samples per each vowel were manually segmented around the vowel onset point at both first and second syllable position of the word. Different signal lengths of the vowel were examined: 5 ms, 10 ms, 20 ms, 30 ms, 40 ms, 50 ms, 60 ms and 70 ms. Each speaker contributed 12 vowel sounds. The total speech sample number for training and testing set was 480 and 240 respectively. The summary was shown in Table 2.

Table 2: Summary of speech database

Set	Features	Number
Training	Speech samples	480
	Speakers	40
	Samples per vowel	80
Testing	Speech samples	240
	Speakers	20
	Samples per vowel	40

The speech tokens were sampled at 20 kHz with 16-bit resolution. The database was collected in normal room environment.

3. SPEECH FEATURE EXTRACTION

A frame of vowel signal was segmented manually around the vowel onset point from each of the Malay words. The vowel signal was preemphasized to flatten the signal.

$$\tilde{s}(n) = s(n) - 0.95s(n-1) \quad (1)$$

The length of the vowel signal was examined from 5 ms to 70 ms. The frame of the vowel signal was then Hamming windowed, to set the signal zero at the beginning and end of the frame.

$$\tilde{x}(n) = x(n)w(n),\ 0 \le \text{n} \le \text{N-1} \quad (2)$$

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (3)$$

The windowed signal was then autocorrelated according to the equation, where the highest autocorrelation value, p is the order of the LPC analysis.

$$R(n) = \sum_{n=0}^{N-1-m} \tilde{x}(n)\tilde{x}(n+m),\ \text{m} = 0, 1, 2, \ldots, \text{p} \quad (4)$$

The selection of p depended primarily on the sampling rate. For a sampling rate of 20 kHz, the value could range from 20 to 24 as suggested in [3].

The autocorrelation coefficients were then converted into Linear-Predictive Coding (LPC) coefficients. Levinson-Durbin recursive algorithm was used perform the conversion.

$$E_0 = R(0) \quad (5)$$

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} a_j^{i-1} R(i-j) \right] / E_{i-1}, 1 \le \text{i} \le \text{p} \quad (6)$$

$$a_i^i = k_i \quad (7)$$

$$a_j^i = a_j^{i-1} - k_i a_{i-j}^{i-1},\ 1 \le \text{j} \le \text{i-1} \quad (8)$$

$$E_i = (1 - k_i^2) E_{i-1} \quad (9)$$

The set of equations (5-9) was solved recursively for i = 1, 2, 3, …, p, where p was the order of the LPC analysis. The k_i were the reflection or PARCOR coefficients. The a_j were the LPC coefficients. The final solution for the LPC coefficients was given as

$$a_j = a_j^{(p)},\ 1 \le \text{j} \le \text{p} \quad (10)$$

The LPC coefficients were converted to cepstral coefficients, which was more robust to noise.

$$c_0 = a_0$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k},\ 1 \le \text{m} \le \text{p} \quad (11)$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k},\ \text{m} > \text{p} \quad (12)$$

The cepstral coefficients were weighted to reduce the sensitivity to noise.

$$\tilde{c}_m = w_m c_m,\ 1 \le \text{m} \le \text{p} \quad (13)$$

$$w_m = \left[1 + \frac{p}{2} \sin\left(\frac{\pi m}{p}\right) \right],\ 1 \le \text{m} \le \text{p} \quad (14)$$

Lastly, the weighted cepstral coefficients were normalized in between +1 and –1 before feeding into the Multi-layer Perceptron.

$$w_{normalized} = 2\left(\frac{w - w_{\min}}{w_{\max} - w_{\min}}\right) - 1 \quad (15)$$

4. MULTI-LAYER PERCEPTRON

A multi-layer perceptron (MLP) with one hidden layer was used to recognize the vowel sounds. The MLP had six output neurons, which corresponded to six Malay vowels.

The MLP was trained to reduce its training errors to a preset minimum value. The E_{rms} was calculated over all the training patterns in an epoch.

$$E_{rms} = \sqrt{\frac{1}{PK}\sum_{p=1}^{P}\sum_{k=1}^{K}(t_{pk} - y_{pk})^2} \qquad (16)$$

The MLP was trained with error back-propagation. The weights were updated after presentation of each training pattern. The sequence of the epoch was randomized every epoch to allow the presentation of the pattern to the input of the network in a random way. This pattern mode of training was preferred because it could avoid the network to be stuck in local minimum, and thus achieving global minimum.

$$w\ (t+1) = w(t) + \Delta w\ (t+1), \qquad (17)$$

where $\Delta w\ (t+1) = \eta\delta\ X + \alpha\Delta w\ (t)$ (18)

In the above equations, η was the learning rate and α was the momentum term. δ was the error correction term at the output layer and hidden layer. X was the vectors to the hidden layer and input layer.

The weights and biases of the MLP were initialized randomly in between –0.3 and +0.3. The target values were set 0.9 and 0.1 to indicate the on and off status respectively. The learning rate and momentum term was set at 0.1 and 0.9 respectively. The MLP was trained to achieve a minimum training error of 0.05 or an epoch of 10000. The MLP was trained and tested with hidden neuron number between 20 and 100, in a step of 10. The size of the input layer depended on the number of cepstral coefficients extracted.

5. RESULTS AND DISCUSSIONS

The performance of the MLP at different cepstral orders and vowel length was shown in Table 3. For each vowel length and cepstral order, the MLP was trained and tested with different hidden neuron number at different training errors. Only the highest recognition accuracy was displayed in the Table 3.

Generally, the performance of the MLP increased with the cepstral order. The cepstral order of 16 showed the poorest accuracy at most of the vowel lengths. Cepstral order of 22 achieved the highest performance at most of the vowel lengths, with an accuracy of 76.25%. Nevertheless, the experimental results showed that cepstral order of 22 to 24 was appropriate for the speech feature extraction at a sampling rate of 20 kHz.

Table 3: Performance of MLP at different signal lengths and cepstral orders

Vowel Length	Cepstral Order				
	16	18	20	22	24
5ms	67.50	67.50	70.00	70.00	71.25
10ms	67.92	69.17	72.92	72.08	72.08
20ms	68.33	70.00	71.67	72.92	72.50
30ms	70.00	70.00	73.75	74.58	75.00
40ms	71.25	71.67	72.92	76.25	74.58
50ms	71.67	70.83	73.75	75.42	74.17
60ms	71.25	70.42	74.58	76.25	75.83
70ms	70.83	71.67	75.00	76.25	76.25

Obviously, the performance of MLP increased as the vowel length increased. The MLP achieved the highest accuracy at several vowel lengths such as 40 ms, 60ms and 70 ms. If average accuracy was calculated over cepstral order 22 to 24, the vowel length of 70 ms surpassed others with an average accuracy of 76.25%. Even though, the average accuracy was calculated over all the cepstral orders, the vowel length of 70 ms still maintained the highest accuracy of 74.00%. The experimental result suggested that longer vowel length was more efficient than short signal length.

The MLP was trained to a preset minimum training error of 0.05. A minimum training error or a maximum training recognition rate did not guarantee a maximum recognition rate of test set. From the Figure 1, the recognition rate of the training set increased from the start of iteration, then moved to the peak, and afterwards fell gradually with the further increment of the recognition rate of training set. Though the recognition rate of the training set is 100% at the minimum training error, but the MLP was only able to achieve a recognition rate of testing set at 69.58%. The MLP achieved the highest recognition rate of test set at training error of 0.105344 or at an iteration of 43, with an accuracy of 76.25%.

The performance of MLP at different hidden neuron number was shown in Table 4. Obviously, the result again showed that the lowest the training error or higher the recognition rate of the training set did not result in higher recognition of test set. The MLP did not generalize well at lower hidden neuron number. At higher hidden neuron number, the MLP was subjected to a bigger network, which did not promise for a better recognition rate.

The confusion matrix of the test set was shown in Table 5. Vowel /i/ was fully recognized by the MLP. The vowel /a/ was the worst to be recognized with an accuracy of 60.00%.

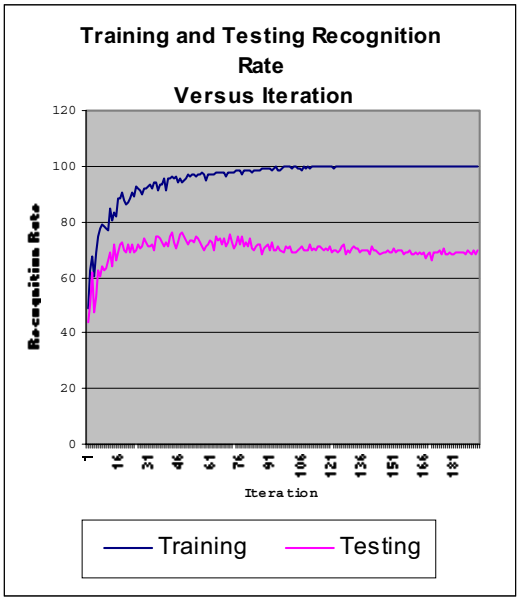

Figure 1: Recognition rate vs iteration

Table 4: Performance of MLP at different hidden neuron numbers

Hidden No	Highest Accuracy of Test Set	Training Error	Iteration	Accuracy of Training Set
20	73.75	0.126	31	92.50
30	74.58	0.114	43	95.00
40	75.00	0.115	37	91.88
50	75.00	0.078	77	98.54
60	76.25	0.105	43	96.46
70	72.92	0.114	37	94.17
80	72.50	0.114	37	92.29
90	72.92	0.118	36	92.92
100	74.17	0.099	55	96.25

Table 5: Confusion matrix of the test set

	/a/	/e/	/ə/	/i/	/o/	/u/	Other	Acc.
/a/	24	1	5	0	4	0	6	60.00
/e/	0	34	1	0	0	0	5	85.00
/ə/	6	3	26	0	1	1	3	65.00
/i/	0	0	0	40	0	0	0	100.00
/o/	2	0	1	0	30	5	2	75.00
/u/	2	0	0	0	4	29	5	72.50

6. CONCLUSION

A speaker-independent Malay vowel recognition using Multi-layer Perceptron was described. The system was able to achieve a recognition rate of 76.25% by just examing a single frame of vowel signal of 70 ms. Experimental results suggested that longer vowel length was preferred over short signal length. Even though the optimal cepstral order was 22 at a sampling rate of 20 kHz, the cepstral order of 22 to 24 was appropriate for extracting the speech features. MLP with the lowest training error or the maximum training recognition rate did not always guarantee the maximum recognition rate of test set.

7. ACKNOWLEDGMENTS

We would like to thank the Ministry of Science, Technology and Innovation (previously Ministry of Science, Technology and Environment), Malaysia for funding the project under IRPA grant. We would like to thank the school for allowing us to collect speech samples for the study.

8. REFERENCES

[1] A. Potamianos, and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603-616, 2003.

[2] M. Giurgiu, "On the use of Neural Networks for automatic vowel recognition," *International IEEE/IAS Conference on Industrial Automation and Control*, pp. 479-484, 1995.

[3] L. Rabiner, and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall Inc, New Jersey, 1993.

[4] A. Hassan, *Linguistik Am Untuk Guru Bahasa Malaysia*, Penerbit Fajar Bakti Sdn. Bhd., Petaling Jaya, 1980.

[5] N. S. Karim, F. M. Onn, H. Musa, and A. H. Mahmood, *Tatabahasa Dewan,* new ed., Dewan Bahasa dan Pustaka, Kuala Lumpur, 1995.

[6] J. Byorick, R. P. Ramachandran, and R. Polikar, "Isolated vowel recognition using linear predictive features and neural networks classifier fusion," *Proceedings of the Fifth International Conference on Information Fusion*, vol. 2, pp. 1565-1572, 2002.